Pattern-based Topic Models for Information Filtering

Yang Gao
Faculty of Science and Engineering
QUT, Brisbane, Australia
Email: y10.gao@student.qut.edu.au

Yue Xu
Faculty of Science and Engineering
QUT, Brisbane, Australia
Email: yue.xu@qut.edu.au

Yuefeng Li
Faculty of Science and Engineering
QUT, Brisbane, Australia
Email: y2.li@.qut.edu.au

***Abstract*—Topic modelling, such as Latent Dirichlet Allocation (LDA), was proposed to generate statistical models to represent multiple topics in a collection of documents, which has been widely utilized in the fields of machine learning and information retrieval, etc. But its effectiveness in information filtering is rarely known. Patterns are always thought to be more representative than single terms for representing documents. In this paper, a novel information filtering model, Pattern-based Topic Model (PBTM) , is proposed to represent the text documents not only using the topic distributions at general level but also using semantic pattern representations at detailed specific level, both of which contribute to the accurate document representation and document relevance ranking. Extensive experiments are conducted to evaluate the effectiveness of PBTM by using the TREC data collection Reuters Corpus Volume 1. The results show that the proposed model achieves outstanding performance.**

***Index Terms*—Topic models, user modelling, pattern mining, closed pattern, information filtering**

I. Introduction

Information filtering (IF) is a system to remove redundant or unwanted information from an information or document stream based on document representations which represent users' interest. Traditional IF models were developed based on a term-based approach, whose advantage is efficient computational performance, as well as mature theories for term weighting, like Rocchio, BM25, et al [1], [2]. But term-based document representation suffers from the problems of polysemy and synonymy. To overcome the limitations of term-based approaches, pattern mining based techniques have been used for information filtering and achieved some improvements on effectiveness [3], [4], since patterns carry more semantic meaning than terms. Also, data mining has developed some techniques (i.e., maximal patterns, closed patterns and master patterns) for removing the redundant and noisy patterns [5]–[8]. One of the promising techniques is Pattern Taxonomy Model (PTM) [9] that discovered closed sequential patterns in text classification. It shows a certain extent improvement on effectiveness, but still faces one challenging issue which is low frequency of the patterns appearing in documents. In order to solve this problem, Wu et.al [10], [11] proposed deploying pattern approach to weight terms by calculating their appearance in discovered patterns.

All these data mining and text mining techniques hold the assumption that user's interest is only related to a single topic. However, the reality is that multiple semantic topics [12] are involved. Topic modelling [13], [14] has become one of the most popular probabilistic text modelling techniques and quickly been accepted by machine learning and text mining communities. The most inspiring contribution of topic modelling is that it automatically classifies documents in a collection by a number of topics and represents every document with multiple topics and their corresponding distribution. Latent Dirichlet Allocation (LDA) [14] is the most effective topic modelling, which has been applied to information retrieval and other application domains and achieved good performance [13], [15]. It is reasonable to expect that applying LDA to IF could make a breakthrough for current IF models due to two advantages of LDA: first, the topic based representation generated by using LDA conquers the problem of semantic confusion compared with the traditional term based document representation. Second, LDA can describe documents at a detailed level with multiple topics instead of a single topic in traditional IF. However, directly applying LDA to IF using topic distributions to represent documents cannot produce satisfactory results due to limited dimensions (i.e., a pre-specified number of topics) in the topic representation, meanwhile, word based topic representation lacks distinguished semantic meaning.

Considering the benefits from data mining, our previous work, called two-stage LDA model, has been proposed in [16] which alleviates the problem of semantic ambiguous topics in LDA by providing a promising way to meaningfully represent topics by patterns rather than single words. However, the pattern-based topic representation generated from the two-stage model represents the collection (of documents) by topics at a high level rather than individual documents. It still remains unsolved that how to utilize the pattern-based topic modelling for document representation. In this paper, we propose a new model called Pattern-based Topic Model (PBTM) to generate document models by using more semantic features based on the pattern based topic representations, and rank the relevance of documents based on user's information needs.

Experiments on evaluating the performance of the proposed PBTM and some baseline models have been conducted on a popular benchmark data collection. The results strongly indicate the outstanding the effectiveness of the proposed model. In Section II, we discuss related work about the state-

of-art IF models. Section IV and V presents the details of our proposed model. Then, we describe data sets, baseline models and experimental results in Section VI. According to the results, we discuss the conclusion and future work in Section VIII.

II. Related Work

IF systems acquire user information needs from "user profiles". IF systems are commonly personalized to support long-term information needs of a particular user or a group of users with similar needs [17]. In IF process, the primary objective is to perform a mapping from a space of incoming documents to a space of user relevant documents. More precisely, denoting the space of incoming documents as D, the mapping $rank : D \rightarrow R$ such that $rank(d)$ corresponds to the relevance of a document d. The relevance of document can be modelled by various approaches that primarily include term-based model, pattern-based model [10], [11], probabilistic model [18], [19] and language model [20], [21].

Data mining techniques were applied to text mining and classification by using word sequences as descriptive phrases (n-Gram) from document collections [22], [23]. But the performance of n-Gram is highly restricted due to low frequency of phrases. Pattern mining has been extensively studied for many years. A variety of efficient algorithms such as Apriori, PrefixSpan, and FP-tree have been proposed and extensively developed for mining frequent patterns more efficiently. But Normally, the number of returned patterns is huge because that if a pattern is frequent, each of its sub-patterns is frequent too. Thus, selecting reliable patterns [8] is always very crucial.

The LDA-based document models [13], [21], [24] are state-of-art topic modelling approaches. Information retrieval systems based on these models achieved good performance. The authors claimed the retrieval performance achieved by [13] not only because of the multiple topics document model, but also because that each topic in the topic model is represented by a group of semantically similar words, which solve the synonymy problem of single words. The relevant documents are determined by user-specific topic model that has been extracted from user information needs [25]. These topic model based applications are all related to long-term user needs extraction and related to the task of this paper. But, there is a lack of explicit discrimination in most of the language models based approaches [26] and probabilistic topic models.

III. Latent Dirichlet Allocation

Topic modelling algorithms are used to discover a set of hidden topics from collections of documents, where a topic is a distribution over terms. Topic models provide an interpretable low-dimensional representation of documents.

Latent Dirichlet Allocation (LDA) [14] is a typical statistical topic modelling technique and the most common topic modelling tool currently in use. It can discover the hidden topics in collections of documents with the appearing words. Let $D = \{d_1, d_2, \cdots, d_M\}$ be a collection of documents. The total number of documents in the collection is *M*. The idea behind LDA is that every document is considered involving multiple topics and each topic can be defined as a distribution over fixed vocabulary of terms that appear in documents. Specifically, LDA models a document as a probabilistic mixture of topics and treats each topic as a probability distribution over words. For the *i*th word in document d, denoted as $w_{d,i}$, the probability of $w_{d,i}$, $P(w_{d,i})$ is defined as:

$$P(w_{d,i}) = \sum_{j=1}^{V} P\left(w_{d,i}|z_{d,i} = Z_j\right) \times P\left(z_{d,i} = Z_j\right) \qquad (1)$$

$z_{d,i}$ is the topic assignment for $w_{d,i}$, $z_{d,i} = Z_j$ means that the word $w_{d,i}$ is assigned to topic j and the V represents the total number of topics. Let ϕ_j be the multinomial distribution over words for Z_j, $\phi_j = (\varphi_{j,1}, \varphi_{j,2}, \cdots, \varphi_{j,n})$, $\sum_{k=1}^{n} \varphi_{j,k} = 1$. θ_d refers to multinomial distribution over topics in document d. $\theta_d = (\vartheta_{d,1}, \vartheta_{d,2}, \cdots, \vartheta_{d,V})$, $\sum_{j=1}^{V} \vartheta_{d,j} = 1$. $\vartheta_{d,j}$ indicates the proportion of topic j in document d. LDA is a generative model in which the only observed variable is $w_{d,i}$, while the others are all latent variables that need to be estimated. Blei, et al. [14] introduce Dirichlet to the posterior probability ϕ_j and θ_d, which optimize the distributions.

Among many available algorithms for estimating hidden variables, the Gibbs sampling method is a very effective strategy for parameter estimation [27] that is used in this paper. After a sufficient number of sampling iterations, the estimated $\hat{\phi_j}$ and $\hat{\theta_d}$ of word-topic distribution and topic-document distribution can be obtained. The resulting representations of LDA are at two levels, collection level and document level. At document level, each document d_i is represented by topic distribution θ_{d_i}. At collection level, D is represented by a set of topics each of which is represented by a probability distribution over words, ϕ_j for topic j. Overall, we have $\Phi = \{\phi_1, \phi_2, \cdots, \phi_V\}$ for all topics. Based on the distribution of Φ for the whole collection, D can be represented by topics distribution, $\theta_D = (\vartheta_{D,1}, \vartheta_{D,2}, \cdots, \vartheta_{D,V})$, $\vartheta_{D,j}$ indicates the proportion of topic j in the collection D.

Apart from these two level outcomes, LDA also generates word-topic assignment, that is, the word occurrence is considered related to the topics by LDA. Take a simple example and let $D = \{d_1, d_2, d_3, d_4\}$ be a small collection of four documents with 12 words appearing in the documents. Assuming the documents in *D* involve 3 topics, Z_1, Z_2 and Z_3. Table I illustrates topic distribution over documents and word-topic assignments in this small collection.

IV. Pattern Based Topic Model

Pattern based representations are considered more meaningful and more accurate to represent topics. Moreover, pattern based representations contain structural information which can reveal the association between terms. In order to discover semantically meaningful and efficient patterns to represent topics and documents, two steps are proposed: firstly, construct a new transactional dataset from the LDA outcomes of the document collection D; secondly, generate pattern based representations from the transactional dataset to represent user needs of the collection D.

TABLE I
EXAMPLE RESULTS OF LDA: WORD-TOPIC ASSIGNMENTS

Topic	Z_1		Z_2		Z_3	
Document	$\vartheta_{d,1}$	words	$\vartheta_{d,2}$	words	$\vartheta_{d,3}$	words
d_1	0.6	w_1, w_2, w_3, w_2, w_1	0.2	w_1, w_9, w_8	0.2	w_7, w_{10}, w_{10}
d_2	0.2	w_2, w_4, w_4	0.5	w_7, w_8, w_1, w_8, w_8	0.3	w_1, w_{11}, w_{12}
d_3	0.3	w_2, w_1, w_7, w_5	0.3	w_7, w_3, w_3, w_2	0.4	w_4, w_7, w_{10}, w_{11}
d_4	0.3	w_2, w_7, w_6	0.4	w_9, w_8, w_1	0.3	w_1, w_{11}, w_{10}

TABLE III
THE FREQUENT PATTERNS FOR Z_2, $\sigma = 2$

Patterns	$supp$
$\{w_1\}, \{w_8\}, \{w_1, w_8\}$	3
$\{w_9\}, \{w_7\}\{w_8, w_9\}, \{w_1, w_9\}, \{w_1, w_8, w_9\}$	2

A. Construct Transactional Dataset

Let R_{d_i,Z_j} represent the word-topic assignment to topic Z_j in document d_i. R_{d_i,Z_j} is a sequence of words assigned to topic Z_j. For the example illustrated in Table I, for topic Z_1 in document d_1, $R_{d_1,Z_1} = \langle w_1, w_2, w_3, w_2, w_1 \rangle$. We construct a set of words from each word-topic assignment R_{d_i,Z_j} instead of using the sequence of words in R_{d_i,Z_j}, because for pattern mining, the frequency of a word within a transaction is insignificant. Let I_{ij} be a set of words which occur in R_{d_i,Z_j}, $I_{ij} = \{w | w \in R_{d_i,Z_j}\}$, i.e., I_{ij} contains the words which are in document d_i and assigned to topic Z_j by LDA. I_{ij}, called a *topical document transaction*, is a set of words without any duplicates. From all the word-topic assignments R_{d_i,Z_j} to Z_j, we can construct a transactional dataset Γ_j. Let $D = \{d_1, \cdots, d_M\}$ be the original document collection, the transactional dataset Γ_j for topic Z_j is defined as $\Gamma_j = \{I_{1j}, I_{2j}, \cdots, I_{Mj}\}$. For the topics in D, we can construct V transactional datasets. An example of transactional datasets is illustrated in Table II, which is generated from the example in Table I.

B. Generate Pattern based Representation

The basic idea of the proposed pattern based method is to use patterns generated from each transactional dataset Γ_j to represent Z_j. In the two-stage topic model [16], frequent patterns are generated in this step. For a given minimal support threshold σ, an itemset X in Γ_j is frequent if $supp(X) >= \sigma$, where $supp(X)$ is the support of X which is the number of transactions in Γ_j that contain X. Take Γ_2 as an example, which is the transactional dataset for Z_2. For a minimal support threshold $\sigma = 2$, all frequent patterns generated from Γ_2 are given in Table III.

V. THE PROPOSED IF MODEL

Representations generated by pattern based LDA carry more concrete and identifiable meaning than the word based representations generated using original LDA. However, the number of patterns in some of the topics can be huge and many of the patterns are not discriminative enough to represent different topics. As a result, documents are hardly accurately represented by these topic representations. That means, these pattern based topic representations which represent user interests may be not sufficient or accurate enough to be directly used to determine the relevance of new documents to the user interests. In this section, a novel IF model based on PBTM is proposed which, instead of directly using the pattern based topic representations, utilizes some semantic patterns including frequent patterns or frequent closed patterns to filter out irrelevant documents.

A. Structural Pattern

For a collection of documents D, by using the pattern based model discussed in Section IV, we can generate the user's interests $U = \{\mathbf{X}_{Z_1}, \mathbf{X}_{Z_2}, \cdots, \mathbf{X}_{Z_V}\}$, and $\mathbf{X}_{Z_i} = \{X_{i1}, X_{i2}, \cdots, X_{im_i}\}$, where $\mathbf{X}_{Z_i}$ is the pattern based representations for Z_i and m_i is the total number of patterns in $\mathbf{X}_{Z_i}$, V is the total number of topics.

Normally, the number of frequent patterns is considerable large and many of them are not necessarily useful. Several concise patterns have been proposed to represent useful patterns generated from a large dataset instead of frequent patterns such as maximal patterns [5] and closed patterns [6]. The number of these concise patterns is significantly smaller than the number of frequent patterns for a dataset. Especially, closed pattern has drawn great attention due to its attractive features [7], [8].

Definition 1. *Closed Itemset*: for a transactional dataset, an itemset X is a closed itemset if there exists no itemset X' such that (1) $X \subset X'$, (2) $supp(X) = supp(X')$.

Closed pattern reveals the relations of the largest range of the associated terms. It covers all information that its subsets describe. Closed patterns are more effective and efficient to represent topics than frequent patterns.

Definition 2. *Generator*: for a transactional dataset Γ, let X be a closed itemset and $T(X)$ consists of all transactions in Γ that contain X, an itemset g is said a generator of X iff $g \subset X, T(g) = T(X)$ and $supp(X) = supp(g)$. A generator g of X is said a minimal generator of X if $\nexists\, g' \subset g$ and g' is a generator of X.

One useful property to be used in the proposed model is pattern taxonomy. Among a set of patterns, usually pattern taxonomy exists. For example, Fig. 1 depicts the taxonomy

TABLE II
TRANSACTIONAL DATASETS GENERATED FROM TABLE I

transaction	topic document transaction	transaction	topic document transaction	transaction	topic document transaction
1	$\{w_1, w_2, w_3\}$	1	$\{w_1, w_8, w_9\}$	1	$\{w_7, w_{10}\}$
2	$\{w_2, w_4\}$	2	$\{w_1, w_7, w_8\}$	2	$\{w_1, w_{11}, w_{12}\}$
3	$\{w_1, w_2, w_5, w_7\}$	3	$\{w_2, w_3, w_7\}$	3	$\{w_4, w_7, w_{10}, w_{11}\}$
4	$\{w_2, w_6, w_7\}$	4	$\{w_1, w_8, w_9\}$	4	$\{w_1, w_{11}, w_{10}\}$
	Γ_1		Γ_2		Γ_3

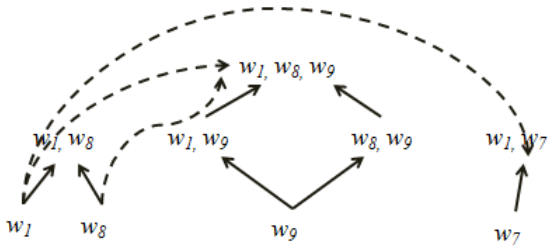

Fig. 1. Pattern Taxonomy in Z_2

constructed for $\mathbf{X}_{Z_2}$ in Table III. This tree-like structure demonstrates the subsumtion relationship between discovered patterns in Z_2. $\{w_1, w_8, w_9\}$ is the most specific pattern in describing user's interests since longer pattern has more specific meaning, while single word, such as w_1 is the most general pattern which is less capable of discriminating the meaning of the topic from other topics than $\{w_1, w_8, w_9\}$. The pattern taxonomy presents different specificities of patterns according to the levels in the taxonomy structure.

As mentioned in pattern taxonomy, the longer the pattern is, the more specific it is. As the result, the specificity of a pattern can be estimated as a function of pattern length. For example, the single word 'mining' usually represents the '-ing' form of 'mine' and it has a general meaning indicating any kind of 'prospecting', whereas 'pattern mining' represents a specific technique in data mining. "Closed pattern mining" is even more specific but still in the same technique area. Generally, the specificity is not necessarily linearly increasing as the pattern size increases. Based on our experimental results, the increase of specificity of a pattern should be slower than the increase of the pattern size. Therefore, we define the pattern specificity as below.

Definition 3. *Pattern specificity:* The specificity of a pattern X is defined as a power function of the pattern length with the exponent less than 1, denoted as $spe(X)$, $spe(X) = a|X|^m$, a and m are constant real numbers and $0 < m < 1$.

B. Document Modelling and Ranking in PBTM

This section presents our novel IF model PBTM, which is based on pattern based LDA. The model consists of two parts, training part to generate user interests from a collection of training documents (i.e., document modelling) and filtering part to determine the relevance of incoming documents based on the user information interests generated in training part (i.e., document ranking).

1) Topic based User Interest Models: By using the methods described in Section IV, for a document collection D and V pre-specified latent topics, from the results of LDA to D, V transactional datasets, $\Gamma_1, \cdots, \Gamma_V$ can be generated from which the pattern based topic representations for the collection, $U = \{\mathbf{X}_{Z_1}, \mathbf{X}_{Z_2}, \cdots, \mathbf{X}_{Z_V}\}$, can be generated, each $\mathbf{X}_{Z_i} = \{X_{i1}, X_{i2}, \cdots, X_{im_i}\}$ is a set of frequent patterns generated from transactional dataset Γ_i. U is considered the user interest model, the patterns in each $\mathbf{X}_{Z_i}$ represent what the user is interested in terms of topic Z_i. Moreover, assume that there are n_i closed patterns in $\mathbf{X}_{Z_i}$, which are $c_{i1}, \cdots, c_{in_i}$, and their corresponding support are $f_{i1}, \cdots, f_{in_i}$, respectively.

2) Topic based Relevance Ranking: According to the literature in Section III, the principle of topic modelling is that user interests or documents are generally represented by topic distribution. But it only represents collection at general level. Therefore, the topic distribution based models cannot represent the collection at semantic level and thus are not be able to distinguish documents which have different semantic focuses. The innovative idea of the proposed model is that, in addition to using topic distribution to represent a collection, the proposed model also represents topics using semantic patterns. We choose two widely used patterns, frequent patterns and closed patterns to represent topics and the pattern support is used to represent topic relevance.

- Frequent pattern based topic model denoted as PBTM_FP
- Closed pattern based topic model denoted as PBTM_FCP

For a new coming document d, the basic idea to determine the relevance of d to the user interests is firstly to identify patterns in d which match some patterns in the user interest model and then estimate the relevance of d by using the support of these patterns and the specificity of the patterns. In order to describe the method, we define the concept of topic significance of a topic to a document.

Definition 4. *Topic Significance*: Let d be a document, Z_j be a topic in the user interest model, PA^d_{jk} be matched patterns,$k = 1, \cdots, n_j$, to document d, and $f_{j1}, \cdots, f_{jn_j}$ be the corresponding supports of the matched patterns within Z_j,

the topic significance of Z_j to d is defined as:

$$sig\left(Z_j, d\right) = \sum_{k=1}^{n_j} spe\left(PA_{jk}^d\right) \times f_{jk} = \sum_{k=1}^{n_j} |PA_{jk}^d|^m \times f_{jk} \quad (2)$$

where m is the scale of pattern specificity, we set $m = 0.5$, a is a constant real number, in this paper, we set $a = 1$.

For a collection of documents D, the user's interests can be represented by the patterns in the topics of D. Furthermore, as discussed in Section III, θ_D represents the topic distribution of D. Therefore, θ_D can be used to represent the user's topic interest distribution. θ_D is topic distribution over a collection documents D, $\theta_D = (\vartheta_{D,1}, \vartheta_{D,2}, \cdots, \vartheta_{D,V})$, $\sum_{j=1}^{V} \vartheta_{D,j} = 1$, V is the number of topics. The user's topic interest distribution can be represented by the collection topic distribution, which is modelled by distribution of topics, which differentiates the degree of topics representing multiple aspects of user interests.

For an incoming document d, we propose to estimate the relevance of d to the user interest based on the topics significance and topics distribution which represents the user's interest to the topic. The document relevance is estimated using the following equation:

$$rank(d) = \sum_{j=1}^{V} sig(Z_j, d) \times \vartheta_{D,j} \quad (3)$$

By incorporating Equation (3), the relevance of d can be estimated by the following equation:

$$rank(d) = \sum_{j=1}^{V} \sum_{k=1}^{n_j} |PA_{jk}^d|^m \times f_{jk} \times \vartheta_{D,j} \quad (4)$$

The higher the $rank(d)$ is, the more likely the document is relevant to the user's interest.

VI. Evaluation

The main hypothesis proposed in this paper is that user information needs involve multiple topics, document modelling by taking multiple topics into consideration can generate more accurate user information needs. To verify this hypothesis, experiments and evaluation have been conducted. This section discusses the experiments and evaluation in terms of data collection, baseline models, measures and results. The results show that, the proposed topic based PBTM model significantly outperforms the state-of-the-art models in terms of effectiveness.

A. Data

In Reuters Corpus Volume 1 (RCV1), there are a total of 806,791 documents that cover a variety of topics and a large amount of information. 100 collections of documents were developed for TREC filtering track. In TREC track, a collection is also referred to as a 'topic'. In this paper, to differentiate from the 'topic' in LDA model, 'collection' is used to refer to a collection of documents in the TREC dataset. The first 50 collections are composed by human assessors and the another 50 collections are constructed artificially from intersections collections. In this paper, only the first 50 collections are used for experiments. For each collection, some documents in RCV1 are divided into a training set and a testing set. According to Buckley and others [28], the 100 collections are stable and enough for high quality experiments. This research uses RCV1 and the 50 assessor collections to evaluate the proposed model. Documents in RCV1 are marked in XML. The 'title' and 'text' of the documents are used by all the models in the experiments.

B. Measures

The effectiveness is assessed by five different measures: average precision of the top K ($K = 20$) documents, $F_\beta(\beta = 1)$ measure, Mean Average Precision (MAP), break-even point (b/p) and Interpolated Average Precision (IAP) on 11-points. F_1 is a criterion that assesses the effect involving both precision (p) and recall (r),which is defined as $F_1 = \frac{2pr}{p+r}$. The larger the *top20*, MAP, b/p or F_1 measure score is, the better the system performs. The 11 points measure is the precisions at 11 standard recall levels (i.e., recall = 0, 0.1,$\cdots$, 1).

The experiments tested cross the 50 collections of independent datasets, which satisfy the generalized cross-validation for statistical estimation model.

C. Baseline Models and Setting

The experiments are conducted extensively to evaluate the effectiveness of the proposed PBTM model, which includes PBTM_FP and PBTM_FCP. The evaluations are conducted in terms of two technical categories: pattern mining methods and topic modelling methods. For each category, some state of the art methods are chosen as the baseline models. The proposed model PBTM is a topic modelling method. For the topic modelling category, the classical LDA method is chosen as the baseline. In addition, we also proposed another topic modelling methods using words (LDA_word) to represent user interests to compare with the proposed model PBTM. For the pattern mining category, the baseline models include frequent closed itemsets (FCP), frequent sequential closed pattern (PTM) and phrases (*n*-Gram).

1) Pattern based category:

- FCP

Frequent closed patterns can more effectively cover the semantics of given dataset than frequent patterns. Moreover, the number of closed patterns is much smaller than that of frequent patterns. Therefore, closed pattern based representation can effectively reduce the size of frequent pattern based representations.

- PTM

The PTM model is the state-of-art pattern based model. It was developed to discover sequential closed patterns from training dataset and ranks the incoming document in filtering stage with the relative supports of discovered patterns that appear in the document.

In PTM model, every document in training dataset (D) is split in paragraphs which are the transactions for pattern

mining. Readers who are interested in the details about PTM are referred to [9] and [11].

- n-Gram

Most researches on phrases in modelling documents have employed an independent collocation discovery module. In this way, a phrase with independent statistics can be indexed exactly as an word based representation. In our experiments, phrases are utilized to represent information needs that are discovered by n-Gram model where $n = 3$.

The minimum support or frequency in pattern-based models, including phrases, sequential closed patterns and frequent closed patterns, is set to 0.2.

2) Topic modelling based category:

- LDA

The baseline LDA [14] classification model directly uses the topic distribution as document representation or user interests. For each of the training document d_i, a topic distribution θ_i can be generated by LDA, i.e., from the training dataset which contains n documents, a set of topic distributions, $Q(\theta) = \{\theta_1, \theta_2, \cdots, \theta_n\}$ can be obtained. In the filtering stage, for every incoming documnent d, we caculate the Kullback-Leibler distance [29] between the topic distribution $\theta_d = (\theta_{d1}, \cdots, \theta_{dV})$, and each $\theta_i = (\theta_{i1}, \cdots, \theta_{iV})$ of the topic distributions in $Q(\theta)$, defined as $KL(d,i) = \sum_{j=1}^{V} \theta_{d,j} \ln \frac{\theta_{dj}}{\theta_{ij}}$, then choose the smallest $KL(d,i)$ as the distance between d and the user's interests, as defined in Equation (5) below:

$$dis(d) = min_{i=1}^{n}(KL(d,i)) = min_{i=1}^{n}\left(\sum_{j=1}^{V} \theta_{dj} \ln \frac{\theta_{dj}}{\theta_{ij}}\right) \quad (5)$$

The smaller the distance $dis(d)$ is, the more likely the document d is relevant to user's interests.

- LDA_word

Words associated with different topics are used to represent user interest needs and word frequency is used to represent topic relevance. The document relevance is calculated by Equation (4). But the specificities of these words equal to 1.

3) Settings: The parameters for both LDA and PBTM are set as follows: the number of iterations of Gibbs sampling is 1000, the hyper-parameters of LDA $\alpha = 50/V, \beta = 0.01$. Our experience shows that filtering results are not very sensitive to the settings of these parameters. But the number of topics V affects the results depending on various data collections. In this paper, V is set to 10.

In the process of generating pattern based topic representations, the relative minimum support σ_{rel} for every topic in each collection is different, because the number of positive documents in collections of RCV1 are very different. In order to ensure enough transactions from positive documents to generate accurate patterns for representing user needs, the minimum support σ_{rel} is set as follows :

$$\sigma_{rel} = \begin{cases} 1 & n \leq 2 \\ max(2/n, 0.3) & 2 < n \leq 10 \\ max(3/n, 0.3) & 10 < n \leq 13 \\ max(4/n, 0.3) & 13 < n \leq 20 \\ 0.3 & otherwise. \end{cases} \quad (6)$$

where n is the number of transactions from relevant documents in each transactional database.

TABLE IV
COMPARISON OF ALL MODELS OVER ALL ASSESSING COLLECTIONS OF RCV1

Methods	*top20*	*b/p*	*MAP*	F_1
PBTM_FCP	**0.494**	**0.420**	0.424	**0.424**
PBTM_FP	0.47	0.402	**0.428**	0.424
LDA_word	0.447	0.410	0.415	0.423
LDA	0.337	0.295	0.308	0.339
$change\%$	+5.1	+4.5	+3.1	+0.23
PTM	0.406	0.353	0.364	0.390
n-Gram	0.401	0.342	0.361	0.386
FCP	0.428	0.346	0.361	0.385
$change\%$	+15.4	+19.0	+16.5	+8.7

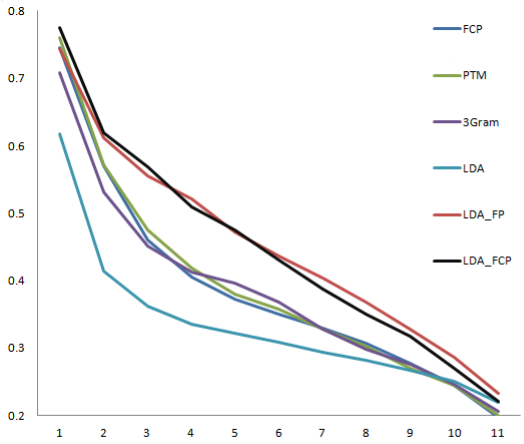

Fig. 2. Comparison between the proposed method and baseline models

D. Results

PBTM_FCP and PBTM_FP are compared with all the baseline models mentioned above, using the 50 human assessed collections. The results evaluated using the measures in Section VI-B. The results are shown in TABLE IV. There are two sections in the table. The top section provides the results of the topic modelling based models and the bottom section provides results for pattern based models. For each section, the change percentage of the proposed model against the best performance for each measure is given in the line labelled with 'change%'.

1) PBTM vs topic-based models: From the top part of Table IV, we can see that, directly using topic distribution from LDA produces disappointing results (i.e., the bottom line in the top part) that are even worse than any other baseline models. The hurting performance of LDA indicates that topic distributions cannot be simply adopted to represent the user's information needs which actually require specific features.

For topic-based models, models with topic based relevance ranking achieves much better performance. Especially, PBTM_FCP outperforms the other models on $top20$ and b/p, while PBTM_FP performs the best on MAP, and they perform the same on F_1. For this category, we can see that PBTM_FCP outperforms LDA_word with a change percentage of 5.1% on $top20$ and 4.5% on b/p, while PBTM_FP outperforms LDA_word with a change percentage of 3.1% on MAP. For F_1, both PBTM_FCP and PBTM_FP outperform LDA_word by a change percentage of 0.23%.

2) PBTM vs Pattern-based Models and n-Gram: We can see that among the three baseline models, PTM outperforms the other two models for b/p, MAP and F_1, while the FCP model performs the best for *top20*. The bottom line of the pattern-based section in the table provides the percentage of improvement achieved by PBTM_FCP against PTM for b/p, MAP and F_1, and against FCP model for *top20*. In the pattern based category, PBTM_FCP achieves excellent performance in improvement percentage with maximum 19.0% and minimum 8.7%, respectively.

The *11-points* results of all methods are shown in Figure 2, which clearly indicates that the proposed model (PBTM) has achieved the best performance comparing with all the other models. Therefore, we can conclude that the experimental results validate the hypothesis that document modelling by taking multiple topics into consideration can model the user information needs more accurately.

Therefore, we conclude that the PBTM is an exciting achievement in discovering high-quality features in text documents.

VII. Discussion

As we can see from the experiment results, taking topics into consideration in generating user interest models and also in document relevance ranking can greatly improve the performance of information filtering. The reasons for achieving the excellent performance of PBTM is mainly because we creatively incorporate pattern mining techniques into topic modelling to generate pattern based topic models which can represent user interest needs in terms of multiple topics. Most importantly, the topic relevance reveals the specificity of topics in detailed level, which brings concrete and precise semantics to document relevance. Moreover, PBTM_FCP in most cases outperforms PBTM_FP since that instead of using all frequent patterns in the user interest model, the concise and quality closed patterns are used to estimate document relevance.

As mentioned in pattern-based baseline models, the transactional datasets for generating patterns usually use sentences or paragraphs as transactions. That an itemset is frequent means it is contained in many paragraphs. It makes sense to some extent when the collection of documents focuses only on one topic. In the case that multiple topics are involved in the collection, the frequent patterns generated from the whole collection may not be able to represent any of the topics and thus hardly to represent the collection.

To emphasize the semantic structure of user's interests which involve multiple topics, PBTM constructs transactional databases in terms of different topics. As the results, transactions in the same topical transactional database share relatively common interest. The discovered patterns from one topical transactional dataset are more likely to represent one aspect of user's interests and more sensitive to get accurate and comprehensive representations of this aspect.

PBTM consists of two parts, topic modelling and pattern mining. For topic modelling that generates user models, the complexity of each iteration of Gibbs sampling for LDA is linear with the number of topics (V) and the number of documents (N), i.e., $O(VN)$ [13].

For pattern mining, no specific quantitative measure for the complexity of pattern mining in relevant literatures. But the efficiency of FP-Tree algorithm has been widely accepted in the field of data mining and text mining. PBTM has the same computational complexity as PTM or frequent closed patterns, on the other hand, PBTM generates patterns from very small transactional datasets comparing with the datasets used in general data mining tasks, because the transactional datasets used in PBTM are generated from the topic representations produced by LDA rather than the original documents. The topic representations containing the words which are considered representing the document topics by LDA are part of the original documents, whereas other pattern mining models generate patterns from the whole collection of documents.

Most importantly, PBTM model combines topic model and pattern mining linearly. Thus in summary, the complexity of PBTM can be determined by topic modelling or pattern mining. In most cases, the complexity of PBTM would be the same as pattern mining since, in general, the complexity of pattern mining is higher than that of topic modelling.

VIII. Conclusion

This paper presents an innovative model PBTM for information filtering including user interest modelling and document relevance ranking. PBTM firstly generates pattern based topic representations to model user's information interests with multiple topics; then PBTM selects quality patterns for estimating the relevance of documents. The proposed approach incorporates the semantic topics from topic modelling and the specificity of the representative patterns. The proposed model has been evaluated by using RCV1 and TREC topics for the task of information filtering. Comparing with the state-of-art models, PBTM demonstrates excellent strength on document modelling and relevance ranking.

The proposed new model automatically generates discriminative and semantic rich representations for modelling topics and documents by combining statistical topic modelling

techniques and data mining techniques. In the future, we can select more discriminative and precise patterns for representing topics and document relevance.

REFERENCES

[1] S. Robertson, H. Zaragoza, and M. Taylor, "Simple bm25 extension to multiple weighted fields," in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*. ACM, 2004, pp. 42–49.

[2] F. Beil, M. Ester, and X. Xu, "Frequent term-based text clustering," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 436–442.

[3] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal, "Mining frequent patterns with counting inference," *ACM SIGKDD Explorations Newsletter*, vol. 2, no. 2, pp. 66–75, 2000.

[4] H. Cheng, X. Yan, J. Han, and C.-W. Hsu, "Discriminative frequent pattern analysis for effective classification," in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 2007, pp. 716–725.

[5] R. J. Bayardo Jr, "Efficiently mining long patterns from databases," in *ACM Sigmod Record*, vol. 27, no. 2. ACM, 1998, pp. 85–93.

[6] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: current status and future directions," *Data Mining and Knowledge Discovery*, vol. 15, no. 1, pp. 55–86, 2007.

[7] M. J. Zaki and C.-J. Hsiao, "Charm: An efficient algorithm for closed itemset mining." in *SDM*, vol. 2, 2002, pp. 457–473.

[8] Y. Xu, Y. Li, and G. Shaw, "Reliable representations for association rules," *Data & Knowledge Engineering*, vol. 70, no. 6, pp. 555–575, 2011.

[9] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic pattern-taxonomy extraction for web mining," in *Web Intelligence, 2004. WI 2004. Proceedings. IEEE/WIC/ACM International Conference on*. IEEE, 2004, pp. 242–248.

[10] S.-T. Wu, Y. Li, and Y. Xu, "Deploying approaches for pattern refinement in text mining," in *Data Mining, 2006. ICDM'06. Sixth International Conference on*. IEEE, 2006, pp. 1157–1161.

[11] N. Zhong, Y. Li, and S.-T. Wu, "Effective pattern discovery for text mining," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 24, no. 1, pp. 30–44, 2012.

[12] E. Cambria, T. Mazzocco, and A. Hussain, "Application of multi-dimensional scaling and artificial neural networks for biologically inspired opinion mining," *Biologically Inspired Cognitive Architectures*, vol. 4, pp. 41–53, 2013.

[13] X. Wei and W. B. Croft, "Lda-based document models for ad-hoc retrieval," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2006, pp. 178–185.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[15] Y. Sun, J. Han, J. Gao, and Y. Yu, "itopicmodel: Information network-integrated topic modeling," in *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*. IEEE, 2009, pp. 493–502.

[16] Y. Gao, Y. Xu, Y. Li, and B. Liu, "A two-stage approach for generating topic models," in *Advances in Knowledge Discovery and Data Mining, 17th Pacific-Asia Conference, PADKDD'13*. Springer, 2013, pp. 221–232.

[17] J. Mostafa, S. Mukhopadhyay, M. Palakal, and W. Lam, "A multilevel approach to intelligent information filtering: model, system, and evaluation," *ACM Transactions on Information Systems (TOIS)*, vol. 15, no. 4, pp. 368–399, 1997.

[18] K. Sparck Jones, S. Walker, and S. E. Robertson, "A probabilistic model of information retrieval: development and comparative experiments: Part 2," *Information Processing & Management*, vol. 36, no. 6, pp. 809–840, 2000.

[19] J. Lafferty and C. Zhai, "Probabilistic relevance models based on document and query generation," in *Language modeling for information retrieval*. Springer, 2003, pp. 1–10.

[20] F. Song and W. B. Croft, "A general language model for information retrieval," in *Proceedings of the eighth international conference on Information and knowledge management*. ACM, 1999, pp. 316–321.

[21] L. Azzopardi, M. Girolami, and C. Van Rijsbergen, "Topic based language models for ad hoc information retrieval," in *Neural Networks, 2004. Proceedings. IEEE International Joint Conference on*, vol. 4. IEEE, 2004, pp. 3281–3286.

[22] J. Fürnkranz, "A study using n-gram features for text categorization," *Austrian Research Institute for Artifical Intelligence*, vol. 3, no. 1998, pp. 1–10, 1998.

[23] W. B. Cavnar, J. M. Trenkle *et al.*, "N-gram-based text categorization," *Ann Arbor MI*, vol. 48113, no. 2, pp. 161–175, 1994.

[24] X. Yi and J. Allan, "A comparative study of utilizing topic models for information retrieval," in *Advances in Information Retrieval*. Springer, 2009, pp. 29–41.

[25] Y. Zhang, J. Callan, and T. Minka, "Novelty and redundancy detection in adaptive filtering," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2002, pp. 81–88.

[26] C. Zhai, "Statistical language models for information retrieval," *Synthesis Lectures on Human Language Technologies*, vol. 1, no. 1, pp. 1–141, 2008.

[27] M. Steyvers and T. Griffiths, "Probabilistic topic models," *Handbook of latent semantic analysis*, vol. 427, no. 7, pp. 424–440, 2007.

[28] C. Buckley and E. M. Voorhees, "Evaluating evaluation measure stability," in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2000, pp. 33–40.

[29] S. Kullback, "The kullback-leibler distance," *The American Statistician*, vol. 41, no. 4, pp. 340–341, 1987.